James Taylor

Assignment 4

University of Maryland University College

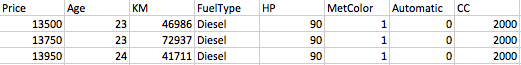
DATA610-9040 Summer 2019

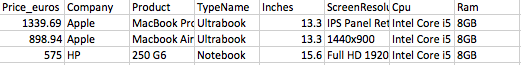
Dr. Saleeb

james.taylor.ps3@gmail.com

**Introduction**

This assignment used two datasets for the purposes of delivering insights from dashboards. They are used car prices and laptop prices.

The used car prices have 1,436 observations. Each observation has the price, number of kilometers the car has driven, the age of the vehicle, the type and size of engine, number of doors, kind of transmission, listed horsepower and color. I set out to understand how an adrenaline junkie (horsepower enthusiast) should look to find the best deal. I did not find obvious outliers or missing data so the set was used as it came. Below show some variables (not all) and a few observations.

The other dataset had 1,303 observations of laptop prices. Every observation has the price, the company that produced it, the model, screen size and resolution, CPU, GPU RAM, Memory size, operating system and weight. I set out to find what drives a laptop’s price. This dataset had outliers I could determine to be wrong so no observations were deducted. Below show some variables (not all) and a few observations.

**Dashboard Presentations**

Just like any purchase, the buyer wants to know if they are getting a good deal. The quality of the deal depends on the products ability to meet the buyer’s expectations. For cars, this means not only long product-life and reliability, but color or feel of the ride. People are willing to pay more for the abstract and subjective pieces of a car if they get enjoyment out of them.

There is undeniably a sect of the population that are captivated by fast cars. The number of movies where the plot revolves around driving fast, and motorsports have been a popular pastime for such a long time speaks to this. I wanted an infographic to help give context to people who want cars with significant horsepower. One can only determine if the deal they are offered is good by comparing it to other deals. The purpose of *Figure 1-Cars* is to give this context. It displays the average price and average kilometers driven for each age of the car. There are strong relationships between the age of a vehicle and its price and kilometers driven. I wouldn’t say there are any surprises here. Older cars have accumulated more miles, so they may be less reliable, less efficient and have less modern technology. A linear regression would be able to determine if age or kilometers driven has more of an effect on price by controlling for both simultaneously, but I didn’t explore this. I would hypothesis that both age and kilometers driven by themselves effect price.

One interesting point is that the rate of price decrease in newer cars is more than the in older. Although it is not a dramatic difference, it appears cars younger than 32 decreases in price less per year than cars older than 32. This seems logical because a car that aged from 20 to 25 increased its age by 25%. Alternatively, a car that ages from 40 to 45 increased it age by 12.5%. Even though both cars aged by 5 years, the younger car aged twice as much relatively speaking.

New cars are more expensive because they are younger and have low miles, but they also have more horsepower. *Figure 2-Cars* show the relation between age and horsepower. It shows a very strong relationship between how old a vehicle is and its horsepower. The size of the circles indicates the average CC’s of the vehicle. That is an acronym for Cubic Centimeters which is a measurement of the displacement an engine has with its pistons. Larger and more pistons means more CC’s, making a larger and more powerful engine. It appears the younger cars have more CC’s than older cars but the increase in CC’s probably doesn’t completely explain the increased horsepower. There were surely technological advances that aided power, like turbochargers or superchargers. This shows our horsepower-enthusiast audience that’s their options lay in younger cars.

Cars have a weak relationship between its price and its horsepower, shown in *Figure 3-Cars.* In the exceptionally high horsepower cars the price does seem to be high. These could be sports cars. In all other cars, it’s not a relationship which is obvious. I wouldn’t imagine horsepower being high on most people’s priorities. Engines that produce more horsepower tend to be less fuel efficient as well. Luxury cars tend to be bigger and heavier, requiring more horsepower to accelerate. Vehicles have more safety and technology now adding weight. All of this noise disrupts the link between horsepower and market price.

The goal of this infographic is to deliver context to a buyer looking for a powerful car. *Figure 1* gives the entire layout of the used car market. It shows there is a tradeoff between the age, the kilometers and the car’s price. It asks the affordability question below to have the buyer figure out where their options lay in the graph. This sets rational expectations about the age and kilometers the buyer can afford. *Figure 2* shows the horsepower and amount of CC’s by age. New cars have more horsepower and more CC’s, but they may not be affordable given the last graph. *Figure 3* shows if the buyer is receiving a good dollar-per-horsepower deal. The relationship is not strong but it still gives context.

The last visualization shows the relationship between acceleration and price. To accelerate fast one must have lots of lots of power with low weight. *Figure 4-Cars* shows heavy and powerful cars are the most expensive on average. The fastest cars are in the bottom-left of the graph. These don’t seem to be too expensive relative to others. The group best fit for speed and acceleration average about $9,000.

The infographic can’t tell someone if their deal is good or bad. It does deliver context to them, seeing if the deal is comparable to others for similar cars. If they really like the car because of the look or an attribute uncaptured by the dataset, the buyer could still be making a wise decision even if it looks like an outlier by our data.

The other dashboard is a storybook designed to discover what drives laptop prices. Initially it looks like the cost of the product is very dependent on brand, shown in *Figure 5-Laptops*. Just like any product brands attempt to stand out in the market. Strategies include cheapest product, quality for the price, and high-end/top of the line performance.

When a decision tree is, displayed in *Figure 6-Laptops,* created the brand isn’t even included while maintaining a 76% predictive strength. The model chose the RAM as the biggest single differentiator for price. It grouped them logically in small, medium and large RAM sized pieces. The small group had 2, 4 and 6 GBs, medium had 8 and 12 GBs, and large had 16, 24, 32 and 64 GBs. It’s obvious the sequential nature the groupings follow. This suggests a valuable correlation. Not only that, but the price averages of those groups were significantly different. Medium RAM laptops are twice as expensive as small ones, and Large RAM are twice as expensive as medium ones. These correlations are not ignorable.

The next visualization shows the amount of each level of RAM and how it effects price. *Figure 7-Laptops* shows the most numerous RAM options are 4GB and 8GB. This allows a buyer more options to get the other specifications they want like CPU or screen resolution. The table on the right is further proof of the RAM and price relationship. For every increase in RAM the prices increase on average.

There is concern that RAM is correlated with price because RAM is also correlated with other expensive parts. For example, if there is a high-performance CPU in the laptop it would make sense to have more RAM to take advantage of it. In other words, more money spent on CPUs means more money should be spent on RAM. In statistics, the CPU would be a confounding variable. It is an exterior variable that is adding to the perceived relationship between two attributes (Statistics How To, n.d.).

One way to test this is to remove RAM from the model and see how it effects the predictive strength. If RAM is really driving the price then removing from the model should drop the predictive ability dramatically. Our original model had a 76% predictive strength with all of the variables, but with RAM unaccounted for the model still manages a 69% prediction strength. Since the model accounts for only 7% less of the variation once RAM isn’t included, I feel we may have a multicollinearity problem. Multicollinearity is when multiple independent variables are significantly related to the predictor in the same way. For example, imagine you are trying to predict height by using two independent features, weight in pounds and weight in kilos. They are different values of weight but they are measuring the same thing. The two weight measurements would have the exact same relationship with height. Including both of these in a model is redundant and create statistical significance issues. In other words, they are not independent and correlated with each other. In summary, RAM is not by itself a great predictor. Laptops that have lots of RAM also have other expensive parts that drive the price as well.

**Impact in Organization**

Dashboards are a great way to organize visualizations into a coherent idea. The ability to create textboxes and arrows on top of visualizations means a presenter doesn’t need to be there to explain what shown. This adds great potential for my current organization because we have the data and the sales team is showing off what we have. Sending sales teams to present to investor firms in different cities and countries is expensive and time consuming. Dashboards like this would help expose our organization to new clients potentially. We can publish fresh looking dashboards articulating interesting findings from our data. The ability to step-through from one graph to the next builds a story and the text box is there to add additional insight that would’ve gone unnoticed if unmentioned. I find people would be inclined to watch a short video more than read a researched paper when visiting the website, so I would imagine a dashboard being displayed in a video would be a good tool to push out. It would display the data we have and our ability to derive useful information from it.

References

Statistics How To. (n.d.). Confounding variable: simple definition and examples. Retrieved from <https://www.statisticshowto.datasciencecentral.com/experimental-design/confounding-variable/>

Appendix A

Figure 1-Cars



Figure 2-Cars

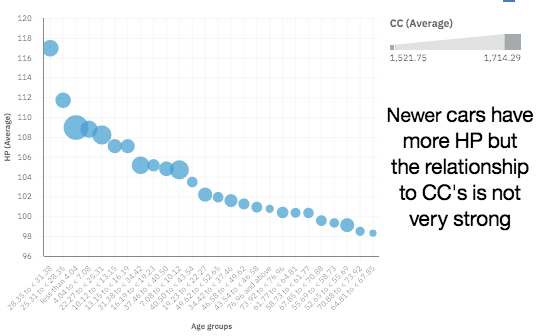


Figure 3-Cars

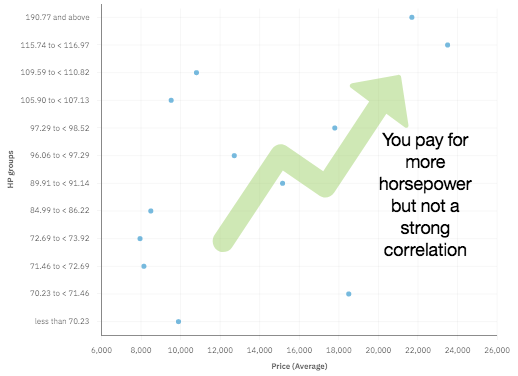


Figure 4-Cars

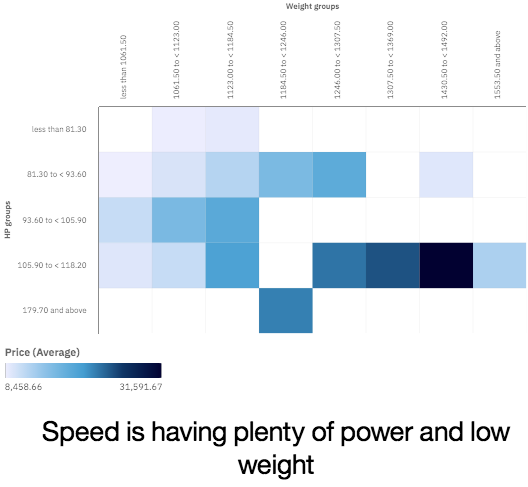


Figure 5-Laptops

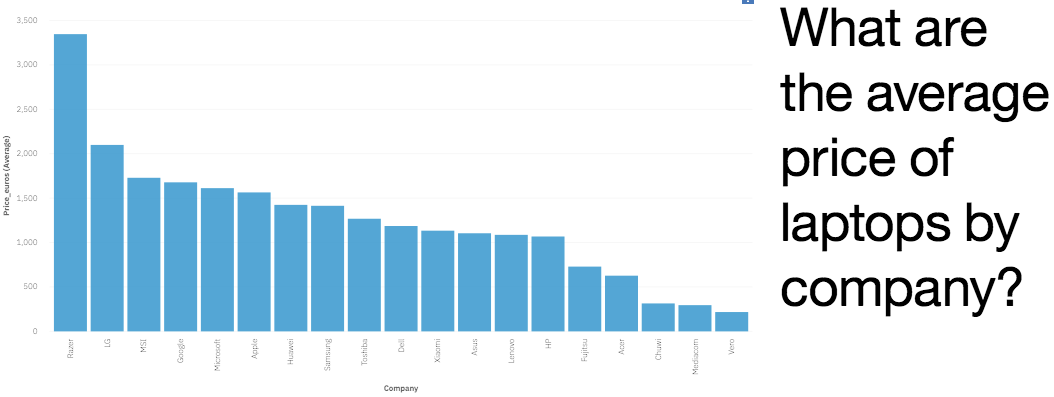


Figure 6-Laptops

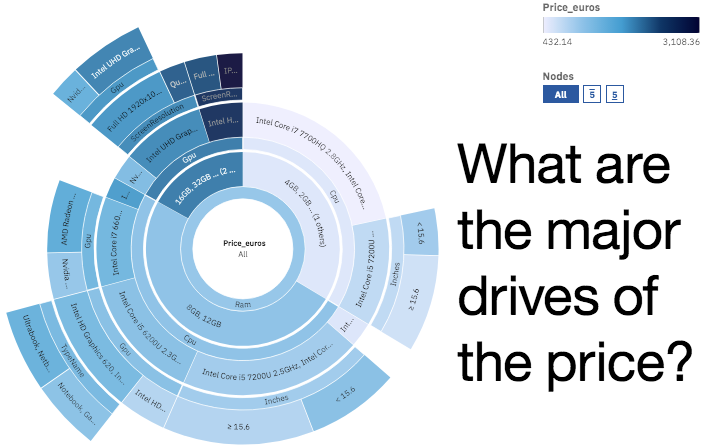


Figure 7-Laptops

